BanglaSenti: A Dataset of Bangla Words for Sentiment Analysis

Hasmot Ali[1], Md. Fahad Hossain[2], Shaon Bhatta Shuvo[3], Ahmed Al Marouf[4]
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
Email: {hasmot15-9632[1], fahad15-9600[2], shaon.cse[3], marouf.cse[4]}@diu.edu.bd

***Abstract*—Being the fifth most spoken language in the world, use of Bangla or Bengali language has spread its breadth into the world of social media. Huge volume of user-generated Bangla data are produced daily in various social media such as Facebook, Twitter, YouTube; online news portals and various websites. Hence, the importance of understanding the emotion and sentiment of these types of data has gain attraction to the researchers' recently. Bangla Natural Language Processing (BNLP) has emerged as a novel research domain because of these multidisciplinary scopes. In this paper, we have presented "*BanglaSenti*", which is a lexicon based corpus or dataset generated solely to identify the sentiment analysis from textual data. This dataset contains 61582 Bangla words with positive, negative and neutral words. These polarities are very significant to understand the overall polarity of the sentences. Not only the corpus, but also a model simulation has been conducted in this paper to understand the usability of this dataset. The dataset is formalized as English SentiWordNet dataset so that researchers' may utilize it with the same format of codes. Though the dataset is developed for sentiment analysis, it could be utilized for emotion detection, opinion/review mining and such applications.**

Keywords—Bangla Natural Langauge Processing (BNLP), Social Media, Bangla Sentiment words, Sentiment Analysis.

I. INTRODUCTION

In the last decade, a dramatic shift in the natural language processing (NLP) research has led to the prevalence of very large-scale applications of statistical methods at the intersection of computer science, artificial intelligence, and linguistics. Natural language processing came into existence to ease the user's work and to satisfy the wish to communicate with the computer in human (natural) language [1]. The input and output of an NLP system can be speech or written Text. Being a dominant research field in NLP, sentiment analysis is the process of determining the opinions of users'. The sentiment can the tracked from various sources, such as comments from social media, or news, blogs, some kind of reviews or opinions. The determination of sentiment not only depends on the polarity, such as positive, negative or neutral; but also depends on the polarity score. Based on these score and labels dataset are formalized to be used for further machine learning algorithms.

Sentiment Analysis aims to classify sentiments from opinions, opinions or attitudes expressed by humans on certain topics of conversation, represented by text. For this purpose, the text can labeled into several categories, for example, positive or negative from the issue sentiment polarity [3]. Sentiment analysis predominantly emphasizes on visions, which express/infer affirmative or undesirable sentiments. The analysis is very useful because it allows us to get an overview of broader public opinions or attitudes towards certain topics, products or services [4].

Almost 200 million people worldwide, 160 million of whom are Bangladeshi speak Bangla as the first language [5]. Bangladeshi people are found to get increasingly involved in online activities such as - expressing their opinions and thoughts on popular microblogging and social networking sites, being connected to friends and families through social media, sharing opinions and thoughts by commenting on online news portals, doing online shopping through online marketplaces and other such applications. Understanding emotion from these user-generated contents has been investigated [14, 15] as it is important to setup business intelligence inside the natural language systems.

Bangla Natural Language Processing (BNLP) has emerged as one of the new research domain and attraction of Bangladeshi research community has embraced this area, as they have heart-felt emotion for the Bangla language. Since few years, the research community is trying to specify the problem areas and the ICT division of the government of Bangladesh also working hard to uphold the research activities. However, one of the commonly faced problem among the community is lack of ground-truth datasets and fair data collection procedures. The dependability and reliability of hand-made datasets are not in satisfactory level for sentiment analysis datasets. Some of the papers tried to create small datasets to work for specific application, which are elaborately discussed in literature review section. Bangla dataset for different purposes such as music stylometric dataset [16] and its applications [17] are found in literature. The scope of this data set is on Bangla music lyrics and its applicability in sentiment analysis, emotion recognition, authorship attribution etc. Therefore, it is going to be challenging to detect subjectivity in huge amounts of online opinions in Bangla language. For extracting sentiment and use them as features in different projects will be time-consuming and challenging. Therefore, we try to help the community to do the same using our ground-truth dataset.

This paper presents the "*BanglaSenti*" dataset which comprises more than forty three thousand words and sentiment polarity and labels along with it. This publicly available dataset could be worked as ground-truth dataset in the application areas of BNLP, such as analysis of sentiment from social media data, detecting depression, analysis of different emotion states etc. The proposed dataset covers the major sentiment words of Bangla language and a model simulation has been conducted in this paper to test the usability of the dataset. With the satisfactory amount and varieties of word collections, this dataset will certainly facilitate the research in BNLP. Rest of the paper cover the related datasets in the literature, dataset creation methodology, and statistical analysis of the dataset, applications and usability criteria of the proposed dataset, simulation of the model dataset and finally the concluding remarks.

II. Literature Review

In past years, sentiment analysis has gained significant attractions in various language perspectives. English, Chinese, Urdu, Bangla and many languages have been used for the same analysis. A good number of dataset is available for sentiment analysis in different language. Analyzing opinion from a language can possible by using the word or sentence of this language. There is a lot of datasets available to analyze sentence such as Twitter [6], Restaurant, Microblog, Cricket, News Comment, Online shopping and many more. However, Baccianella in [7] introduces a new type of dataset called SentiWordNet, which analyzes every Word from a Sentence for Opinion Mining. In 2016 SemEval workshop, a multilingual dataset [9] was introduced in 8 different languages. The dataset was domain specific and covered seven different domains such as laptop, digital camera, restaurant, mobiles, museums, hotels and restaurants [8]. A very few numbers of Dataset for Bangla Sentiment Analysis is available nowadays. Among them, Rahman [10] offers a pair of Dataset involving some comment of Cricket and Restaurant reviews. Mahmudun et al. [11] used the first and last letter of a word to analyze the features and detect sentiment from Bangla text. In [11], presented feature analysis and sentiment analysis techniques using their self-created dataset which contains 1500 short comments and the proposed technique gives approximately 83% accuracy using their own dataset. To analyze Bangla Microblog post Chowdhury [12] used a Dataset from querying Twitter API v1.1 and translate them to Bangla. Most of the Bangla dataset are performing in a specific field of mining. Here the dataset can perform all the task in a single package and analyze Opinion Mining in every field of Bangla Language.

N. Banik et al. [18] presented a comparative analysis of machine learning algorithms on Bangla emotional text analysis. They used two different dataset. First one named Parts-of-Speech (POS) Tagset containing 3,000 Bangla sentences, 42,000 words and 32 tagset which is available in web paid downloads. Another one is their self-developed dataset containing 6314 Facebook comments. They trained 4700 data, test 940 of data and acquired highest accuracy of 52.98%. K. M. Hasan et al [19] tried to adopt contextual valency analysis for detecting sentiment using WordNet and SentiWordNet 3.0 English dataset and found accuracy of about 95%. R. A. Tuhin et al. [20] proposed an automated system to detect sentiment using their own dataset consisting of 7500 Bangla sentences and reported accuracy of about 90%.

Utilizing microblog posts [12] or comments [21-22] for detecting sentiment is found in literature. In [21], S. Chowdhury et al. tried to utilize microblog posts from Twitter and collected 1300 posts and developed machine learning algorithms to identify sentiment. They used 1000 sentence as training data and 300 sentence as test data and attained highest accuracy of 93% by SVM.

In recent days deep learning methods are proven to be better than the traditional machine learning systems in image processing, decision support system as natural language processing as well. Therefore, researchers' have tried to use the deep learning methods such as deep recurrent model [22-23] and convolutional neural network [24]. In [22], they used their own dataset containing 9337 post, where 72% are in Bangla text and another 28% is in romanized Bangla text. They collect 4621 post from Facebook, 2610 post from Twitter, 801 posts from YouTube, 1255 posts from Online News and rest 50 from Product Review Page. In [23], they came up with 80% accuracy on character level supervised learning method and 77% accuracy from baseline model with word level representation using recurrent model. In [24], M. H. Alam et al. presented convolutional neural network based model to detect sentiment from Bangla sentences. They collect approximately to 120000 comment where 50% of the comments are of positive polarity and the model gives around 99% accuracy. But, the models internal descriptions and explanation of this high accuracy are not clearly discussed in the paper. Again, the dataset is not publicly available.

In the literature some extraordinary algorithms or frameworks have been presented such as lexicon based backtracking algorithm [25], multilabel sentiment and emotion detection [26] and empirical framework [27]. In [25], Tapasy et al, presented a backtracking algorithm to detect sentiment from Bangla songs. They have taken 201 Bengali comments from a YouTube channel owned by Bengali new young star Mahtim Sakib. They trained their algorithm with 288 unique expression. In [26], N. I. Tripto et al. has collected comments from different types of video domains using YouTube API version 3.0. They highest achievable accuracy for 3 and 5 class sentiment analysis is 65.97% and 54.24% respectively. For five-category emotion detection, the accuracy is 59.23%. In [27], N. Tabassum et al. has designed an empirical framework to detect sentiment, making a dataset of 1050 Bangla comments from Facebook and twitter. But, the dataset is not publicly available.

The dataset titled as "*BanglaSenti*" is presented in this paper, which will be publicly available for such applications discussed above and the comparison of different machine learning as well as deep learning methods can be performed upon this dataset.

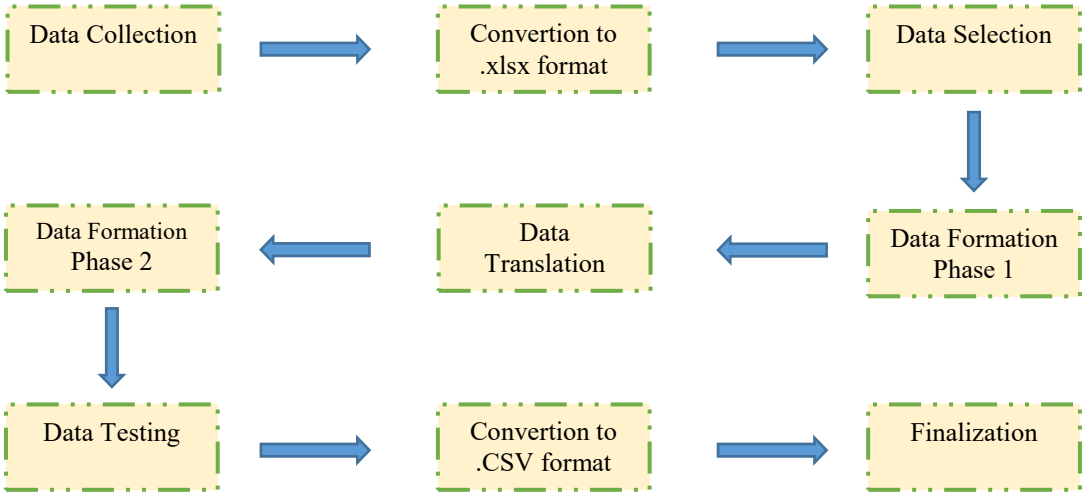

Fig. 1. Dataset creation methodology.

III. Methodology

In this paper, we have adopted several procedural steps to collect, select, format and translate the Bangla words to create our proposed dataset. Our dataset is mostly influenced by SentiWordNet 3.0 [13]. *BanglaSenti* is a collection of 61582 Bangla word and their corresponding score and English translation available in this dataset.

A. Data Collection

We collect the data from SentiWordNet and this is the most updated version of their dataset named SentiWordNet 3.0 which is developed from SentiWordNet 1.0. SentiWordNet 3.0 has a huge collection of 117660 words with corresponding Positive and Negative score as well as unique word id, representative parts of speech, SynsetTerms and Gloss. We collect the data from GitHub in text (.txt) format and convert them to Microsoft Excel Spreadsheets (.xlsx) format for further processing.

B. Data Selection

For preparing a dataset suitable for Bangla language we have to perform several processes to select the right data. SentiWordNet version 3.0 devised parts-of-speech (POS) and SynsetTerms, introducing the sense numbers for the repeated English words. Because of the language perspective of Bangla words, we have deleted the repeated, meaningless (in terms of Bangla language) and the word which is containing Noun. We also have deleted the column of ID, SynsetTerms, and Gloss. Therefore, the selection process was sophisticated in nature and the

C. Data Formating

We split the word containing Number (#) and Numeric value and delete them. Then we focus on the word containing underscore (_). Basically, we wanted to provide a dataset containing only Single word. So we have to replace the underscore with space () and split them in a new row with the corresponding score. And then we again perform the deletion of repeated word. Finally, we format an English dataset of Opinion Mining containing only a single word with Positive and Negative scores for every corresponding word. And then storing the improved data for further processing.

D. Data Translation

As we are going to perform a Bangla sentiment analysis, therefore we have prepare a dataset for Bangla language. Now we are going to translate the preprocessed English version of dataset into Bangla. For performing translation tools we used Microsoft Excel 2016 which actually translates from one to another language using Google Translator using GOOGLETRANSLATE (SourceCellAddress, “EN”, “BN”) function. For doing this we have to place the whole sheet into Google Sheet and collect the translated result with the corresponding score as text format and place into Excel for acquiring translated result as a text, not as a function. Then we perform deletion for deleting repeated Bangla word from the translated Dataset. Then, we have sorted the dataset in Bangla alphabetical order and finalizing by testing with some random Bangla word. For implementing the testing phase, we have used *python* and the recognition rate is quite high. For this paper, we did not run any machine learning system to analyze the accuracy values.

IV. Statistical Analysis of BanglaSenti

In this section, we have presented the statistical analysis of *BanglaSenti*. The quantitative values are presented in the following Table I.

TABLE I. Quantitative Data of BanglaSenti

Properties	Values
Total No. of Words	61582
No. of Positive Words	7443
No. of Negative Words	8274
No. of Neutral Words	45865

TABLE II. PERCENTAGE OF SENTIMENT SCORE IN BANGLASENTI

Score	Positive	Negative
0	83.896959%	83.6667511%
0.125	6.0705817%	4.910334%
0.222	0.0207187%	0.0046%
0.25	3.7523884%	3.7869196%
0.3	0.0046%	0.0046%
0.333	0.0138125%	2.3458183%
0.364	0.0023%	0.0161146%
0.375	2.5299846%	2.2261102%
0.444	0.0230208%	0.0230208%
0.5	2.016621%	1.6782154%
0.556	0.0161146%	0.0023%
0.625	1.1280186%	0.0138125%
0.667	0.0046%	0.0046%
0.75	0.349916%	1.0175188%
0.778	0.0046%	0.0207187%
0.875	0.1542393%	0.2578328%
1	0.0115104%	0.0207187%

V. CONCLUSION

This paper presents "BanglaSenti" dataset, a structured lexicon based corpus to detect sentiment of Bangla words. Total 61582 Bangla words are acquired with the sentiment scores along with the word meanings in English. Several applications regarding BNLP can utilize this dataset. This dataset could be considered as ground-truth dataset for the related area and referred to be the first dataset to be publicly available [28].

REFERENCES

[1] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh, "Natural Language Processing: State of The Art, Current Trends and Challenges", 2017.

[2] B. Liu, "Sentiment Analysis and Opinion Mining". Morgan & Claypool Publishers, 2012.

[3] Deborah Kurniawati, Edy Prayitno, Dini Fakta Sari, Septian Narsa Putra, "Sentiment Analysis of Use on Twitter on Police Institution Services Using Naïve Bayes Classifier Method", Proceedings of the 3rd International Conference of Project Management (ICPM), Bali, 2019.

[4] N. Boudad, R. Faizi, H.O.R. Thami, and R. Chiheb, "Sentiment analysis in Arabic: A review of the literature", AIN Shams Engineering Journal vol. 9, 2479-2490, 2018.

[5] Banglapedia. Bangla Language. [2016 August 30] [Online] http://en.banglapedia.org/index.php?title=Bangla_Language.

[6] Kashfia Sailunaz, Reda Alhajj, "Emotion and sentiment analysis from Twitter Text", Journal of Computational Science, Vol. 36, September, 2019. https://doi.org/10.1016/j.jocs.2019.05.009

[7] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", Proceedings of LREC. 10, 2010.

[8] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, & de clercq, Orphee & Hoste, Veronique & Apidianaki, Marianna & Tannier, Xavier & Loukachevitch, Natalia & Kotelnikov, Evgeny & Bel, Nuria & Zafra, Salud María & Eryiğit, Gülşen, "SemEval-2016 Task 5: Aspect Based Sentiment Analysis". Pp. 19-30. 10.18653/v1/S16-1002, 2016.

[9] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, Manandhar, S. AL-Smadi, M., Al-Ayyoub, M.; Zhao, Y., Qin, B., and De Clercq, O. "SemEval-2016 Task 5: Aspect Based Sentiment Analysis. Available online: http://www.aclweb.org/anthology/S16-1002

[10] Md. Rahman, and E. Dey, (2018). Datasets for Aspect-Based Sentiment Analysis in Bangla and Its Baseline Evaluation. Data.2018 3. 15. 10.3390/data3020015.

[11] M. Mahmudun, M. Tanzir, and S. Ismail, "Detecting Sentiment from Bangla Text using Machine Learning Technique and Feature Analysis". International Journal of Computer Applications (IJCA), vol. 153, pp. 28-34, 2016. 10.5120/ijca2016912230.

[12] S. Chowdhury, and W. Chowdhury, "Performing sentiment analysis in Bangla microblog posts" International Conference on Informatics, Electronics and Vision, ICIEV 2014, pp. 1-6. doi: 10.1109/ICIEV.2014.6850712.

[13] A. Esuli, and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining", 2006.

[14] A. A. Marouf, R, Hossain, R. Sarker, "Understanding Emotional and Language Tone from Music Lyrics using IBM Watson Tone Analyzer", 2019 3rd IEEE International Conference on Electrical, Computer, Communication Technologies (ICECCT 2019), Coimbatore, India, 20-22 February, 2019.

[15] R. Hossain, A. A. Marouf, R. Sarker, M. Mimo, and B.Pandey, "Title Recommendation Approach from English Song Lyrics using Topic Modeling Algorithm", 3rd IEEE International Conference on Electrical, Computer, Communication Technologies (ICECCT 2019), Coimbatore, India, 20-22 February, 2019.

[16] R. Hossain, and A. A. Marouf, "BanglaMusicStylo: A Stylometric Dataset of Bangla Music Lyrics", 1st IEEE International Conference on Bangla Speech and Language Processing (ICBSLP), SUST, 21-22 September, 2018.

[17] A. A. Marouf, and R. Hossain, "Lyricist Identification using Stylometric Features utilizing BanglaMusicStylo Dataset", 2nd IEEE International Conference on Bangla Speech and Language Processing (ICBSLP), SUST, 27-28 September, 2019.

[18] N. Banik, and M. Hasan Hafizur Rahman, "Evaluation of Naïve Bayes and Support Vector Machines on Bangla Textual Movie Reviews," 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, 2018, pp. 1-6.

[19] K. M. Hasan, M. Rahman, and Badiuzzaman, "Sentiment detection from Bangla text using contextual valency analysis", 2014 17th International Conference on Computer and Information Technology, ICCIT 2014.

[20] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter and A. K. Das, "An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, pp. 360-364, 2019,.

[21] S. Chowdhury, and W. Chowdhury, "Performing sentiment analysis in Bangla microblog posts," 2014 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, 2014.

[22] A. Hassan, "Sentiment analysis on Bangla and Romanized Bangla Text using deep recurrent models." 2016 International Workshop on Computational Intelligence (IWCI) (2016), pp. 51-56

[23] M. S. Haydar, M. Al Helal and S. A. Hossain, "Sentiment Extraction From Bangla Text : A Character Level Supervised Recurrent Neural Network Approach," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, 2018, pp. 1-4.

[24] M. H. Alam, M. Rahoman and M. A. K. Azad, "Sentiment analysis for Bangla sentences using convolutional neural network," 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, 2017, pp. 1-6.

[25] T. Rabeya, N. R. Chakraborty, S. Ferdous, M. Dash and A. Al Marouf, "Sentiment Analysis of Bangla Song Review- A Lexicon Based Backtracking Approach," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2019, pp. 1-7.

[26] N. Irtiza Tripto and M. Eunus Ali, "Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments", 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, 2018, pp. 1-6.

[27] N. Tabassum, M. Khan, "Design an Empirical Framework for Sentiment Analysis from Bangla Text using Machine Learning", pp. 1-5, 2019. 10.1109/ECACE.2019.8679347.

[28] Github Link: [Online], https://github.com/fahad35/BanglaSenti-A-Dataset-of-Bangla-Words-for-Sentiment-Analysis